

## Software Implementation of Synchronous Memory Barriers

### BACKGROUND OF THE INVENTION

#### Technical Field

This invention relates to software for implementing synchronous memory barriers in a multiprocessor computing environment. More specifically, the invention relates to a method and system for selectively emulating sequential consistency in a shared memory computing environment.

#### Description Of The Prior Art

Multiprocessor systems contain multiple processors (also referred to herein as "CPUs") that can execute multiple processes or multiple threads within a single process simultaneously in a manner known as parallel computing. In general, multiprocessor systems execute multiple processes or threads faster than conventional single processor systems, such as personal computers, that execute programs sequentially. The actual performance advantage is a function of a number of factors, including the degree to which parts of a multithreaded process and/or multiple distinct processes can be executed in parallel and the architecture of the particular multiprocessor system. The degree to which processes can be executed in parallel depends, in part, on the extent to which they compete for exclusive access to shared memory resources.

Shared memory multiprocessor systems offer a common physical memory address space that all processors can access. Multiple processes therein, or multiple threads within a process, can communicate through shared variables in memory which allow the processes to read or write to the same memory location in the computer system. Message passing multiprocessor systems, in contrast to shared memory system, have a separate memory space for each processor. They require processes to communicate through explicit messages to each other.

A significant issue in the design of multiprocessor systems is process synchronization. The degree to which processes can be executed in parallel depends in part on the extent to which they compete for exclusive access to shared memory resources. For example, if two processes A and B are executing in parallel, process B might have to wait for process A to write a value to a buffer before process B can access it. Otherwise, a race condition could occur, where process B might access the buffer while process A was part way through updating the buffer. To avoid conflicts, synchronization mechanisms are provided to control the order of process execution. These mechanisms include mutual exclusion locks, condition variables, counting semaphores, and reader-writer locks. A mutual exclusion lock allows only the processor holding the lock to execute an associated action. When a processor requests a mutual exclusion lock, it is granted to that processor exclusively. Other processors desiring the lock must wait until the processor with the lock releases it. To address the buffer scenario described above, both processes would request the mutual exclusion lock before executing further. Whichever process first acquires the lock then updates (in case of process A) or accesses (in case of process B) the buffer. The other processor must wait until the first processor finishes and releases the lock. In this way, the lock guarantees that process B sees consistent information, even if processors running in parallel execute processes A and B.

For processes to be synchronized, instructions requiring exclusive access can be grouped into a critical section and associated with a lock. When a process is executing instructions in its critical section, a mutual exclusion lock guarantees no other processes are executing the same instructions. This is important where processors are attempting to change data. However, such a lock has the drawback in that it prohibits multiple processes from simultaneously executing instructions that only allow the processes to read data. A reader-writer lock, in contrast, allows multiple reading processes ("readers") to access simultaneously a shared resource such as a database, while a writing process ("writer") must have exclusive access to the database before performing any updates for consistency. A practical example of a situation appropriate for a reader-writer lock is a TCP/IP routing structure with many readers and an occasional update of the information. Recent implementations of reader-writer locks are described by Mellor-Crummey and Scott (MCS) in "Scalable Reader-Writer Synchronization for Shared-Memory

Multiprocessors,” *Proceedings of the Third ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 106-113 (1991) and Hseih and Weihl in “Scalable Reader-Writer Locks for Parallel Systems,” *Technical Report MIT/LCS/TR-521* (November 1991).

5           The basic mechanics and structure of reader-writer locks are well known. In a typical lock, multiple readers may acquire the lock, but only if there are no active writers. Conversely, a writer may acquire the lock only if there are no active readers or another writer. When a reader releases the lock, it takes no action unless it is the last active reader, if so, it grants the lock to the next waiting writer.

0323456789  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200

Read-copy update is one example of a technique that does not require readers to acquire locks. Another example where readers do not acquire locks is with algorithms that rely on a strong memory consistency model such as a sequentially consistent memory model. Sequentially consistent memory requires that the result of any execution be the same as if the accesses executed by each processor were kept in order and the accesses among different processors were interleaved. One way to implement sequential consistency is to delay the completion of some memory access. Accordingly, sequentially consistent memory is generally inefficient.

Figs. 1 a-c outline the prior art process of adding a new element 30 to a data structure 5 in a sequentially consistent memory model. Fig. 1a is an illustration of a sequentially consistent memory model for a data structure prior to adding or initializing a new element 30 to the data structure 5. The data structure 5 includes a first element 10 and a second element 20. Both the first and second elements 10 and 20, respectively, have three fields 12, 14 and 15, and 22, 24 and 26. In order to add a new element 30 to the data structure 5 such that the CPUs in the multiprocessor environment could concurrently search the data structure, the new element 30 must first be initialized. This ensures that CPUs searching the linked data structure do not see fields in the new element filled with corrupted data. Following initialization of the new element's 30 fields 32, 34 and 36, the new element may be added to the data structure 5. Fig. 1b is an illustration of the new element 30 following initialization of each of its fields 32, 34 and 36, and prior to adding the new element 30 to the data structure 5. Finally, Fig. 1c illustrates the addition of the third element to the data structure following the initialization of the fields 32, 24 and 36. Accordingly, in a sequentially consistent memory model execution of each step in the process must occur in a program order.

The process of Figs. 1a-c is only effective on CPUs that use a strong memory consistency model such as sequential consistency. For example, the addition of a new element may fail in weaker memory models where other CPUs may see write operations from a given CPU happening in different orders. Fig. 2 is an illustration of a prior art weak memory-consistency model for adding a new element to a data structure. In this example, the write operation to the new element's 30 first field 32 passes the write operation to the second element's 20 next field. A CPU searching

the data structure may see the first field 32 of the third element 30, resulting in corrupted data. The searching CPU may then attempt to use the data ascertained from the field 32 as a pointer, and most likely this would result in a program failure or a system crash. Accordingly, data corruption can be avoided by using CPUs that enforce stronger memory consistency.

5 Stronger hardware memory consistency requires more overhead and it cannot implicitly differentiate priority read and write requests. To overcome this problem, modern microprocessors implement relaxed memory consistency models where memory operations can appear to occur in different orders on different CPUs. For example, the DEC/Compaq Alpha has a memory barrier that serializes writes and invalidations, but only with respect to the CPU executing the memory barrier. There is no hardware mechanism to invalidate a data item from all other CPU's caches and to wait until these invalidations are complete. Accordingly, it is desirable to provide a high priority interprocessor interrupt to request that all CPUs in the system execute a memory barrier instruction, thereby requiring both reading and updating CPUs to have passed through a memory barrier to ensure a consistent view of memory.

## SUMMARY OF THE INVENTION

It is therefore an object of the invention to provide software for implementing synchronous memory barriers in a multiprocessor computer system. It is a further object of the invention to process memory invalidates through the use of memory barrier instructions to ensure a consistent view of memory.

20 A first aspect of the invention is a method of selectively emulating sequential consistency in software. Each CPU in the multiprocessing computer environment is forced to execute a memory barrier instruction. Following execution of the memory barrier, each CPU sends an indicator to communicate completion of the memory barrier instruction. An interprocessor interrupt is sent to each CPU to force execution of the memory barrier instruction. To avoid  
25 deadlock, execution of memory barrier instructions from a responding CPU are registered with the

requesting CPU. Furthermore, CPUs waiting for other CPUs to execute memory barrier instructions must remain sensitive to concurrent requests. Implementation of registration of memory barrier instructions is preferably, but not necessarily, selected from the group consisting of a bitmask, an array and a combining tree.

5 A second aspect of the invention is a multiprocessor computer system which includes an instruction for forcing each CPU to execute a memory barrier instruction, and an instruction manager for indicating completion of the memory barrier instruction. A memory barrier manager is provided to send an interprocessor interrupt to all of the CPUs to force execution of the memory barrier instruction. To avoid deadlock between competing CPUs, registration of execution of a memory barrier instruction from a responding CPU is provided to the requesting CPU. In addition, the requesting CPU remains sensitive to and executes concurrent requests. Implementation of registration of instructions is preferably, but not necessarily, selected from the group consisting of a bitmask, an array, and a combining tree.

10 A third aspect of the invention is an article comprising a computer-readable signal bearing medium with multiple processors operating within the medium. The article includes means in the medium for forcing each CPU to execute a memory barrier instruction, and an instruction manager for indicating completion of the memory barrier instruction. A memory barrier manager is provided to send an interprocessor interrupt to all of the CPUs to force execution of the memory barrier instruction. To avoid deadlock between competing CPUs, registration of execution of memory barrier instruction from a responding CPU is provided to the requesting CPU. In addition, the requesting CPU remains sensitive to and executes concurrent requests. Implementation of registration of instructions is preferably, but not necessarily, selected from the group consisting of a bitmask, an array, and a combining tree.

20 Other features and advantages of this invention will become apparent from the following detailed description of the presently preferred embodiment of the invention, taken in conjunction with the accompanying drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1a is a block diagram of a prior art data structure at an initial state.

FIG. 1b is a block diagram of a prior art data structure with a new element initialized.

FIG. 1c is a block diagram of a prior art data structure with a new element appended to a

list.

FIG. 2 is a block diagram of a prior art data structure of a weak memory-consistency model.

FIG. 3 is a block diagram of the interconnect between a reading CPU and a writing CPU.

FIG. 4 is a flow chart illustrating sequential consistency according to the preferred embodiment of this invention, and is suggested for printing on the first page of the issued patent.

## DESCRIPTION OF THE PREFERRED EMBODIMENT

### Overview

In shared memory multiprocessor systems, it is essential that multiple processors see a consistent view of memory. Examples of techniques in which the reading CPU does not acquire a lock is seen in read-copy update and in the implementation of a weak memory barrier instruction in a sequential consistency model. The following is pseudocode for updating a pointer to a new element in a linked list:

1. Initialize new structure, including the pointer to the next element.
2. Register a read-copy callback that invokes a function that awakens all processes sleeping on a semaphore "s".
3. Sleep on semaphore "s".
4. Make the intended predecessor element point to the new element.

Since the read-copy callback implementation forces a memory barrier to execute on each CPU, all reading CPUs are guaranteed to execute a memory barrier between the time that the structure is initialized and the pointer is modified, as is required to avoid memory corruption. However, this procedure is not usable from within interrupt handlers, spinlock critical sections, or sections of code with interrupts disabled. CPUs that use relaxed memory consistency models fail to provide a mechanism to resolve the reading CPUs from reading data that is in the process of being invalidated. Accordingly, it is desirable and efficient to implement a software to emulate sequential consistency while avoiding a deadlock scenario between multiple CPUs.

### Technical Background

In general, implementing a software approach to a weak memory consistency model improves efficiency while alleviating overhead. Fig. 3 refers to a block diagram 50 of a writing CPU 60 and a reading CPU 70 for illustrating the issue of the weak memory barrier execution. Each CPU has even numbered cache lines processed by cache<sub>0</sub> 62 and 64, respectively, and odd numbered cache lines processed by cache<sub>1</sub> 72 and 74, respectively. The cache-line size is 72 bytes with the even numbered cache line at addresses 0, 64, 128, 192, 256..., and the odd numbered cache lines at addresses 32, 96, 160, 224, 288... In this example, the writing CPU 60 is adding a new data structure to the end of a linked list while the reading CPU 70 is concurrently scanning the same linked list. The writing CPU 60 can first fill in the data structure and then update the pointer from the last element in the linked list to point to the new element. However, this procedure can result in the reading CPU 70 seeing garbage values in this new element since the update to the pointer might propagate to the reading CPU faster than the changes to the data structure. For example, the pointer might be in an even numbered cache line and the new element might reside in an odd numbered cache line, and the even numbered hardware might be idle while the odd numbered hardware might be busy. Accordingly, there is a need to implement proper execution of a memory barrier to ensure memory consistency with both CPUs.

To ensure memory consistency with the example illustrated in Fig. 3, the writing CPU 60 must execute a memory barrier instruction after filling in the data structure and before updating the



pointer. In the case where the memory barrier instruction's effect is limited to a single CPU, the memory barrier instruction executed by the writing CPU 60 would force the invalidations to appear in order on the interconnect, but would not guarantee that the reading CPU 70 would process the invalidations in order. For example, the even numbered cache hardware in the reading CPU 70 might be idle while the odd numbered cache hardware on the reading CPU might be busy. This would result in the reading CPU 70 reading the new value of the pointer but seeing the old data in the data element. Accordingly, there is a need to ensure that the old data is invalidated prior to the reading CPU 70 accessing the data structure.

Fig. 4 is a flowchart 100 illustrating a process for updating a pointer to a data structure utilizing the implementation of memory barriers. A CPU updating a data structure may write to the data structure 110. This CPU is referred to as the writing CPU. During this process, each CPU accessing and reading the data structure is utilizing the old data structure until such time as the writing CPU updates the pointer to the new element. Each CPU reading the data structure is known as a reading CPU. Following the update of the data structure, the writing CPU forces each CPU in the system to execute a memory barrier instruction 120. The writing CPU uses a high priority inter-processor interrupt to force all CPUs to execute the memory barrier instruction. The execution of the memory barrier instruction invalidates the stale data and ensures that the reading CPUs will access the new data in the modified data structure. Each CPU sends an indicator to a memory location to indicate completion of the memory barrier instruction 130. This step ensures a recordation in a common location to indicate completion of the memory barrier instruction. Accordingly, the writing CPU forces execution of a memory barrier instruction by each CPU in the system in order to invalidate old data prior to updating the pointer to the data structure for the new data.

However, it is critical to avoid deadlock between two competing writing CPUs. Each writing CPU must ensure that each CPU has completed the memory barrier execution 140 prior to updating the pointer to the data structure. If each CPU has not executed a memory barrier instruction, the writing CPU cannot update the pointer to the data structure. The writing CPU must wait for each CPU to execute a memory barrier instruction, or it may again request execution of a

memory barrier instruction 120. While the writing CPU is waiting for each CPU to register execution of the memory barrier, the writing CPU checks for and satisfies concurrent memory barrier execution requests, 145. When each CPU has registered completion of execution of the memory barrier instruction, the writing CPU may update the pointer to the data structure 150.

5 Accordingly, following a review of the register to ensure that each CPU in the system has executed the memory barrier instruction, the writing CPU may update the pointer to the data structure.

There are two components for avoiding deadlocks, registering completion of execution of memory barrier instructions, and satisfying concurrent memory barrier requests. A first embodiment for avoiding a deadlock scenario is to provide an array for each CPU

10 to register memory barrier execution requests. The array provides for one entry per CPU. A CPU requesting a lock to write to the data structure must scan the array to guarantee that all CPUs have executed a memory barrier to flush out any invalidates prior to updating the pointer. In a preferred embodiment, each entry in the array is a bitmask, with one bit per CPU. The CPU requesting the lock to update the pointer sends an interrupt to all CPUs to force a memory barrier execution.

15 Each CPU uses an atomic operation to subtract its bits from each CPUs bitmask from the array, and the requesting CPU must scan the array until each of the values in the array is zero. During this time, the writing CPU checks for and satisfies concurrent memory barrier requests. This guarantees that each CPU has executed a memory barrier instruction.

Pseudocode for a CPU responding to a request to execute a memory barrier instruction in

20 conjunction with an array entry system of the first embodiment is as follows:

1. Set a local variable "cleared\_bits" to zero.
  2. Suppress interrupts.
  3. Acquire the "need\_mb\_lock".
  4. Scan the "need\_mb" array. For each entry that has this CPU's bit set, do the
- 25 following:
- a) Clear this CPU's bit.
  - b) Increment the "cleared\_bits" local variable.

5. If “cleared\_bits” is non-zero, execute a memory-barrier instruction.
6. Release the “need\_mb\_lock”.
7. Restore interrupts.

Accordingly, step 4 entails the step for registering completion of memory barrier execution.

5 Some architectures allow for combining the acquisition and release of the lock with the suppressing and restoring of interrupts, respectively. Pseudocode for a CPU requesting a global memory-barrier shutdown in conjunction with the array entry system of the first embodiment is as follows:

1. Suppress interrupts.
2. Acquire the “need\_mb\_lock”.
3. Execute a memory-barrier instruction, which may be implied by the acquisition of the lock.
4. Within the “need\_mb\_entry” for this CPU, set the bits for all the other CPUs.
5. Release the “need\_mb\_lock”.
6. Send interrupts to every other CPU.
7. While this CPU’s “need\_mb” entry is non-zero, repeat the following steps:
  - a) Set a local variable “cleared\_bits” to zero.
  - b) Scan the “need\_mb” array. For each entry “j”:
    - i) If the j’th entry has this CPU’s bit set:
      - 20 (1) Acquire the “need\_mb\_lock”.
      - (2) Clear this CPU’s bit.
      - (3) Increment the “cleared\_bits” local variable.
      - (4) Release the “need\_mb\_lock”.
    - c) If “cleared\_bits” is non-zero, execute a memory-barrier instruction.
  - 25 8. Restore interrupts.

Step 7 and the subordinate steps avoid deadlock by causing waiting CPUs to respond to concurrent requests by other CPUs. Accordingly, both of the implementations of the array system require the requesting CPU to scan the array to ensure that each of the other CPUs have executed the memory barrier prior to updating a pointer in the data structure, while requiring waiting CPUs to respond to concurrent memory barrier requests, thereby avoiding deadlock.

A second embodiment for avoiding a deadlock scenario is to use a generation-based bitmask. Each memory-barrier execution request is assigned a generation number. Requests that are initiated while a previous request is being serviced is assigned the same generation number as the previous request. Once a generation number is serviced, a request is complete. Deadlock is avoided by having all waiting CPUs repeatedly execute memory-barrier instructions and registering completion of the memory-barrier instruction with the bitmask.

Pseudocode for a CPU responding to an interrupt requesting a memory-barrier execution implementing the generation based bitmask is as follows:

1. Suppress interrupts.
2. If our bit in "need\_mb" (bitmask) is not set, restore interrupts and return.
3. Acquire "need\_mb\_lock".
4. Execute a memory-barrier instruction
5. Clear out bit in "need\_mb". If ours is the last bit set, do the following:
  - a) Increment the "current generation" counter.
  - b) If the "current generation" counter is less than or equal to the "maximum generation" counter, do the following:
    - i) Set each CPU's bit (except for this CPU's bit) in the "need\_mb" bitmask.
    - ii) Send interrupts to each other CPU.
6. Release "need\_mb\_lock".

Accordingly, step 5 entails the step for registering completion of memory barrier execution.

Pseudocode for a CPU requesting a global memory-barrier shutdown is as follows:

1. Suppress interrupts and acquire “need\_mb\_lock”.
2. Execute a memory-barrier instruction.
3. If the current generation is less than or equal to the maximum generation, do the following:
  - a) Set “maximum generation” to “current generation” + 1.
4. Otherwise, do the following:
  - a) Set “maximum generation” to “current generation”.
  - b) Set “my generation” to “maximum generation”.
  - c) Set each CPU’s bit (except for this CPU’s bit) in the “need\_mb” bitmask.
  - d) Send interrupts to each other CPU.
5. Set “my generation” to “maximum generation”.
6. Release “need\_mb\_lock”.
7. While “my generation” is greater than or equal to “current generation”:
  - a) Invoke the procedure that responds to a request for each CPU to execute a memory barrier instruction.

Accordingly, Step 7 ensures that the waiting CPUs remain sensitive to and respond to concurrent memory barrier requests.

In implementing either the first or second embodiments to avoid deadlock, it is critical to have each reading CPU execute a memory barrier instruction. However, it may be desirable for each CPU, including the writing CPU requesting the memory barrier instruction for the remaining CPUs, to execute a memory barrier instruction. Execution of the memory barrier instruction invalidates old data from each CPU’s cache. Implementation of the memory barrier execution by each CPU is conducted by sending a high priority interprocessor interrupt to all CPUs in the system. This forces each CPU to execute the associated memory barrier instruction. Concurrent memory barrier execution requests are merged into groups wherein each group of requests are assigned a generation number. A current generation number is assigned to all arriving memory

barrier execution requests, while a previous memory barrier execution request is being serviced. The software for emulating sequential consistency further requires each CPU waiting for other CPUs to execute a set of memory barrier instructions to continuously satisfy concurrent memory barrier execution requests. In a multiprocessing computing environment, each CPU may have a  
5 different agenda, and multiple memory barrier execution requests may be processing within a short time interval. Accordingly, it is critical that concurrent memory barrier instruction requests are met to ensure that invalidated data is not consumed in an inappropriate manner.

### Advantages Over The Prior Art

10 The implementation of software for synchronizing memory barrier instructions improves performance in reading and writing operations. By implementing a memory barrier request from the writing CPU, invalidates are forced to be processed before the reading CPUs read the pointer to the data structure. This ensures that the reading CPUs are not reading invalidated data or data that is in the process of being invalidated. A strong hardware consistency model uses more time and therefore has a greater overhead. The implementation of the memory barrier flushes out the  
15 invalidations. Furthermore, hardware cannot implicitly differentiate priority between reading and writing operations. Accordingly, the method for emulating sequential consistency in software reduces costs and ensures avoidance of a deadlock between multiple CPUs.

### Alternative Embodiments

20 It will be appreciated that, although specific embodiments of the invention have been described herein for purposes of illustration, various modifications may be made without departing from the spirit and scope of the invention. In particular, alternative mechanisms may be employed in order to avoid deadlock between multiple CPUs. For example, a set of request counters may be used in conjunction with registers to ensure that no more than one CPU is acting at  
25 any one time. Another mechanism is the use of a combining tree to ensure that no more than one CPU is acting at any one time. A combining tree mechanism may be desirable in a NUMA environment, or in an alternative computer system where it may be desirable to reflect the bus

structure of the computer system. Accordingly, the scope of protection of this invention is limited only by the following claims and their equivalents.

FIG. 10 is a block diagram of the computer system.